Autonomous Generation of Soundscapes using Unstructured Sound Databases

Nathaniel Finney

MASTER THESIS UPF / 2009
Master in Sound and Music Computing

Master thesis supervisor:
Jordi Janer
Department of Information and Communication Technologies
Universitat Pompeu Fabra, Barcelona

.

Abstract

This research focuses on the generation of soundscapes using unstructured sound databases for the sonification of virtual environments. A generalized methodology for design based on soundscape categorization, perceptual discrimination of sources and media design principles is proposed, with the underlying principle of the composition of a source and a textural layer within any soundscape. A generative model is proposed based on these principles covering sound object retrieval, segmentation, parameterization and resynthesis. The model incorporates wavelet resynthesis, sample playback and a technique for concatenative synthesis using an MFCC-based BIC segmentation method. Principles for optimal grain size selection with respect to source layer content are discussed, and the concatenation of segments is based on a relative MFCC Euclidean distance calculation.

An implementation of the model using a photorealistic panoramic image in an urban context is described, using a sound database of community-provided recordings. The implementation utilizes sample playback and concatenative synthesis in order to maximally preserve the contextual attributes of the photorealistic environment, while wavelet resynthesis is discussed as a potential avenue for further development. The methods of classification, segmentation and synthesis adhering to the particular application are discussed, along with a validation of the model using a subjective evaluation. The results of the study demonstrate the applicability of the design principles to an autonomous generation engine, while highlighting some of the challenges of implementation for autonomous functionality related to retrieval, segmentation and synthesis parameterization.

.

Acknowledgments

I would like to sincerely thank my supervisor Jordi Janer for his contributions to this research, including ideas, discussions and technical input. Also, my fellow members of the MTG Metaverse project deserve special appreciation for their efforts: Stefan Kersten for contribution of knowledge of the field and enough literature to fill a library, as well as lessons in Supercollider, and Gerard Roma for invaluable assistance with programming audio in Actionscript, while responding very kindly to my naive programming-related questions, and setting up servers to work around endless security policy file problems.

I also give a large token of appreciation to Xavier Serra for guidance since the beginning of the Master's program in interpreting my entangled interests toward research work that was somehow very in tune with all of them, and in the end an extremely rewarding experience. Furthermore to Emilia Gomez for her coordination efforts of the program, and always showing a genuine interest in the students' progress and contentedness.

I also extend a sincere *thanks* to Hendrik Purwins for contribution of ideas, stimulating discussions that never ceased to be interesting, and sense of humor in the students' heaviest times of stress.

And to the fellow students I have had the privilege to have as colleagues this year, I give my deepest gratitude for sharing this experience and giving your sometimes brutally honest opinions of my ideas and beliefs for research in SMC. It has been an absolute pleasure to work alongside you all, and I look forward to continued work and/or socializing with you in the future.

I would like to personally thank my flatmates over this past year, the members of the *Rock Fort*: Leonardo "*The Tortilla Sensei*" Aldrey, Sašo "*Hide Your Terasse Plantholders!*" Muševič, Sabine "*Chippie*" Mehlin, Olivier "*Tac Tac*" Lalonde, Katalin "*Makes Me Look Good in Pictures*" Karólyi and Robin "*Texan Turf Battle*" Motheral for your friendship, great meals and general support.

And last but far from least, to my family and friends abroad, thank you for continuing to lend me your support and guidance with only the best of intentions.

Contents

Chapter 1

Introduction

The focus of this study is on the application of unstructured sound databases to the autonomous design and generation of soundscapes for virtual environments. The following chapters discuss aspects related to soundscape and sound source classification, sound scene design for media, the perceptual sense of presence in sonic environments, and an implementation of a soundscape generator using unstructured audio databases. An autonomous generation model may be separated into three major portions: retrieval, design and generation. The following work focuses primarily on the design and generation aspects, while implementation of autonomous retrieval is reserved as an avenue for future work.

The design of sonic environments requires a fundamental classification strategy for both soundscapes and sound sources. The classification of soundscapes is most widely performed using R. Murray Schafer's referential taxonomy, which incorporates socio-cultural attributes and ecological acoustics [14]. Many models for analyzing and generating soundscapes use Schafer's breakdown of the components of the soundscape into *keynote* and *signal* layers as a foundation for their modeling of the soundscape [4], [21]. In addition to a referential classification, a sound source is often classified according to its physical characteristics, such as in the work of William Gaver where he developed a taxonomy according to the interaction of materials [10]. The use of such a taxonomy is pinnacle in developing methods for modeling and synthesizing sound objects according to physical attributes, and is a groundwork for many studies in ecological acoustics and physical modeling synthesis [13]. The variability of physical attributes for resynthesizing sound objects is a growing field of research for improving the interactivity aspect in VE's [7]. This study however focuses on the textural and passive elements of the soundscape to which interactive sonic elements may be added.

In the work of Birchfield et. al., the generation of the soundscape is based on a probabilistic selection of audio tracks separated according the Schafer concept of keynote sounds and signal sounds, where the material has been retrieved prior to generation with a lexical search using WordNet [8]. Birchfield et. al. incorporates dynamically changing probabilities in response to user behavior for triggering audio tracks, which results in an evolving sound environment that they propose reflects the sonic diversity of the equivalent natural environment [4]. Probabilistic models for soundscape generation are a demonstration of the notion that a generated sound-

scape is greatly enhanced by variety and evolution, for which research has shown to be a contributing factor to the sense of presence in VE's [16], [20].

With the use of unstructured databases as the audio input to the soundscape generator, such a model may be considered both a framework for community interaction and an ecological acoustics preservational tool, while taking advantage of the growing databases of sonic material such as The Freesound Project [1]. An autonomous soundscape generation model that can incorporate the past and current research in soundscape design and composition would be a tool that would allow for the usage of the increasing audio resources that are at the community's disposal, and for the improvement of immersive qualities within VE's.

Chapter 2

Soundscape Description and Perception

This section describes the most accepted theories and techniques of soundscape classification and design. The design of sonic virtual environments (VE's) utilizes many sound design concepts from film and video games as well as some more recent concepts that have emerged along with the establishment of virtual reality and environments on the internet for social collaboration and networking. Many of the concepts related to immersion in film and video games and the role of the sonic environment are applicable for optimizing immersion in interactive virtual environments, while many aspects unique to the nature of VE's pose new challenges to sound design theories and technologies.

2.1 Soundscape Classification

2.1.1 Schafer's Taxonomy

R. Murray Schafer distinguishes three main themes of a soundscape: keynote sounds, signals and soundmarks [14]. The keynote is analogous to the musical notion of *key* which is the tonal reference point, or *anchor* of a composition. Schafer compares the concept of keynote sounds in soundscapes to the visual perception of figure and ground, where the figure is the subject of attention and the ground is the surrounding context which gives the figure its shape and mass. Without the ground, the figure becomes shapeless and nonexistent [14]. Signals are sounds in the foreground that draw the listener's atttention, and can thus be thought of as the figure in the visual perception analogy. Since theoretically any sound can become a signal sound if the listener decides to pay attention to it, Schafer further defines the signal sound as those sounds that force attention from the listener, such as bells, horns and whistles. Soundmarks are those sounds that are particularly noticeable for a distinct community, deriving from the term landmark.

Schafer's breakdown of the main themes of the soundscape serves as a good starting point for defining and classifying a soundscape. He gives examples of how these themes have changed throughout history, taking examples from rural life and cities, the onset of industrialization and electricity, and how these factors may have caused

drastic changes to the sonic environment in which man has lived. These concepts may be of interest in VE's where environmental preservation is a consideration.

2.1.2 Physical Characteristics

Classification of sounds based on physical characteristics such as material and type of interaction provides a framework for categorizing sound objects, which can be utilized for sonification purposes. In Gaver's taxonomy [10] a hierarchy of materials and their sonic interactions are described, which provides a groundwork for organizing databases of sounds and segmented audio events.

Another classification scheme for individual sound sources based on the physical characteristics was proposed by Pierre Shaeffer, and further elaborated upon by R. Murray Schafer, which categorizes a signal using the quantities attack, body and decay in a tabular format against the quantities duration, mass, grain and dynamics [14]. This method is more generalized than Gaver's taxonomy, and more related to the temporal audio content rather than the physical content.

A method of classification of sound objects which combines the two taxonomies could be very interesting for categorization and retrieval of segmented sound objects. The development of such a scheme is discussed as an avenue for future work in section 5.

2.1.3 Referential Characteristics

In soundscape generation it is very useful to employ a referential classification of sound sources, such as that described by R. Murray Schafer, which groups objects based on their context rather than content or physical characteristics. Schafer divides the sources within soundscapes into six referential categories listed as follows:

1. Natural Sounds
2. Human Sounds
3. Sounds and Society
4. Mechanical Sounds
5. Quiet and Silence
6. Sounds as Indicators

These categories are further subdivided into smaller groups such as churches and fireworks. The entire list is given in Appendix A.

2.2 The Sound Scene in Media

2.2.1 Diegesis

In sound design for film, the sound scene can be divided into two main portions: diegetic and non-diegetic. Diegetic refers to any sound source that can be resolved to a component of the environment, such as foley effects, ambient sounds such as wind and traffic, and any music played by a device or person within the environment. Non-diegetic is any sound source that is not within the environment such as a narrator,

ambience effects not heard by the characters in a story, or music not linked to a source in the environment [5].

Sonnenschein describes the use of non-diegetic sounds in cinema as guiding the listener toward a feeling, subjectively beyond the visual elements giving the examples of a wolf howl piercing a downtown redlight district and an exaggerated ticking clock over an office worker under the pressure of a deadline [19]. It is evident that these types of cinematic effects can have a dramatic effect on a cinematic storyline, and the use of such effects in a VE can be assumed to have an equally dramatic effect if suitably applied in a given context.

For the purposes of this study, the soundscape can be assumed to only be composed of diegetic sounds, and a further implementation regarding the enhancement of the scene can include a layer with sonic additions such as musical accompaniment and additional effects. This constraint implies that all sounds generated in the virtual soundscape must contain a direct relationship to a source either visible by the user or based upon a logical assumption of its source within the environment.

2.2.2 Synchrony

Sound events are furthermore labeled as either synchronous or non-synchronous [5], meaning that their sonic behavior is related to the visual aspects of particular objects within the VE. Some examples of synchrony are spatial position relative to the user, contacts between objects, speech and animal noises, and acoustical effects related to the interaction of the environment and sound sources.

2.2.3 On-Screen and Off-Screen

In cinema the sounds that are related to sources within the frame are termed as on-screen, while those that are not within the frame are called off-screen [19]. Off-screen sounds are further divided into the categories active and passive. Active sounds are those that invoke attention from the subject, and can be considered analogous to Schafer's classification term signals. This distinction is highly influential on the sound design of the scene, as they shape the attention and focus of the subject. However, it is intuitive that the distinction is not discrete, since some sounds may affect the attention of the listener more than others. This level of attractiveness can depend on many properties such as volume, spectral content, termporal fluctuation, and the adhesion with the context.

2.3 The Sense of Presence

This section describes the elements that affect the sense of presence in VE's. The sense of presence is defined as the feeling of being situated in an environment despite being physically situated in another [17]. In quantifying and/or qualifying the sense of presence, or the similar term immersion, different authors have proposed and evaluated various components that make up this highly subjective sense [17], [20], [16].

In order to experience some degree of presence in a VE, the psychological states involvement and immersion are necessary precursors. Involvement is defined as the focus on a coherent set of stimuli, and immersion is the actual perception of envelopment in an environment [17].

Singer, et al. breaks down the psychological factors affecting the sense of presence into four categories: control factors, sensory factors, distraction factors and realism factors. In table 2.1 the four categories with their relative parameters are presented.

Table 2.1: Psychological factors affecting the sense of presence in virtual environments [17].

Control Factors	Sensory Factors	Distraction Factors	Realism Factors
Degree of Control	Sensory Modality	Isolation	Scene Realism
Immediacy of Control	Environmental Richness	Selective Attention	Information Consistent with Obj. World
Anticipation of Events	Multimodal Presentation	Interface Awareness	Meaningfulness of Experience
Mode of Control	Consistency of Multimodal Information		Separation Anxiety and Disorder
Physical Environment Modifiability	Degree of Movement Perception		
	Active Search		

In a study where photorealistic environments with generated soundscapes were presented to subjects, a set of seven criteria were used to evaluate the quality of the soundscape generation for the image [16]. Images of parks and cityscapes were used and the user was asked afterwards to identify the sources within the environments. The authors assumed the following factors to be of principal importance to the subjects' sense of place.

- Sound Delivery Method
- Motion
- Interactivity
- Exaggeration
- Variety
- Quantity

The sound delivery method refers to the manner of presentation of the soundscape, such as headphones versus a surround speaker configuration. Motion of

sources within the environment are propsed by the authors to add to the dynamism of the soundscape, and thereby induce more interest in the user to the environment. Interactivity in this particular study refers to the use of physical models to alter the nature of the sound samples according the position of the user relative to the source, where an example may be footsteps on different materials. The processing of audio content to fit the context is very related to the parameters scene realism, consistency of multimodal information and information consistent with the observed world, as shown in table 2.1.

Exaggeration is a quality that refers to the use of sound effects to enhance the natural character of sound sources for a purpose, such as is used often in cinema and video games. This concept is counteractive to the concepts described above related to realism and consistency, and exposes the compromise between soundscape design and realism. While added elements which may enhance the perceptual response or encourage a certain emotional response may contribute to the immersive qualities of a soundscape, in some applications they may have a distracting effect, which detracts from the cogency of the environment. Therefore it is evident that a soundscape design application requires an initial decision regarding whether the primary intention is realism or emotional response.

Variety is thought to be of critical importance to the perceptual believability of the soundscape, and refers to the change and evolution of the soundscape over time, or in other words, the absence of the recognition of looping sounds. A limitation of the use of the samples for playback is that the user may perceive loops within the soundscape, which may highly reduce the sense of presence in the environment.

Quality refers to the density of sources within the soundscape, and is discussed by the authors in terms of the sources that capture the attention of the subject. If the density of recognizable sources is too high, then the subject may have difficulty focusing, while if it is too low, then the soundscape may cease to be engaging.

Chapter 3

Design and Generation

Using the principles explained in the previous chapter, an outline for optimizing the sense of presence in VE's is described in this section for the application of autonomous soundscape generation. The sound design elements and principles are applied in relation to the factors affecting presence in VE's and the most relevant methods of audio synthesis for soundscape generation.

3.1 Design Methodology

The principles of sound design and soundscape characterization as described in the previous chapter lead to the underlying hypothesis of the design methodology described in this section, that the treatment of the soundscape in terms of two layers, textures (background) and objects (foreground), establishes a framework for algorithmic design and interaction principles more suitable for information delivery and perceptual comfort than with a unified montage of sound sources.

The textural elements and object elements are segmented from sound files retrieved from a database, and categorized into one of these two layers using their semantic identifiers, which are either extracted from tags or a recognition model. Objects are those sounds that are meant or expected to draw attention from the user, and may include indicators, soundmarks or informational content such as church bells or non-diegetic sounds such as narration. Textural elements are determined to be those sources that form the ambience such as birds and wind, and tend to be more stochastic in nature while drawing minimal focus from the user.

Figure 3.1 shows the overview of the system functionality, where the two sound groups, *textures* and *objects* are seen to be handled separately until the final mixing and spatialization. Solid arrows signify transfer of audio content, while dashed arrows are informational streams supplied to the various function blocks. The following subsections contain short descriptions of each block while the finer details of the generation method are described in sections 3.2 and 3.3.

3.1.1 Input Content

The inputs to the system are in the middle of the diagram, signified by circles, and divided into four categories: audio content, locality attributes, real-time conditions

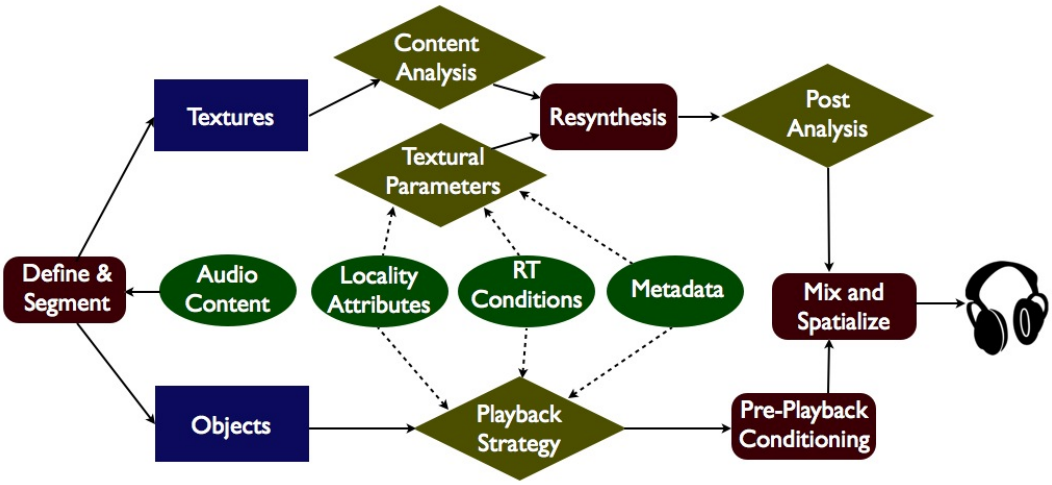

Figure 3.1: Overview of the system functionality with given inputs (green circles), analysis blocks (diamonds), sound groups (rectangles) and processes (rectangles with soft corners).

and metadata.

Audio Content Audio content is retrieved from a database of user-provided audio recordings, and contains all tag information and any accompanying descriptor information related to audio analyses on the server.

Locality Attributes Locality-related details include information such as spoken language, common animal life, vehicle types, and both the density and types of construction machinery. This information is predetermined and applied during the search within the database of sounds for use in the soundscape generation.

The soundmarks for the city are determined using a search based on a predefined list of potential soundmarks. This list includes items such as churches, stadiums and above-ground trains or trolleys. The signals for the city include locality-specific police sirens and crosswalk signals, and are found through a preemptive search for city information.

Real-time conditions The generated soundscape is used to convey information to the user regarding the real-time conditions of the locality such as traffic, weather (including seasonal information), and other current activity taking place at that locality such as demonstrations or festivals.

Metadata Information ascertained from the community and user interaction may provide avenues for optimizing and enriching the soundscape generation. This topic will be discussed further in section 5 as an opportunity for extension to the application.

3.1.2 Analysis

Content Analysis Analysis of the textural sounds is used to obtain spectral and temporal attributes for use in the resynthesis process, and is performed using

basic audio analysis to extract the duration and level, and more involved techniques such as fourier analysis and wavelet analysis for use in the resynthesis process.

Textural Parameters The informational inputs provide a data set that is analyzed to return a set of parameters for the textural resynthesis process, including factors related to the locality-specific attributes, conditions and metadata. These may include characteristics such as the density of people or animals and weather conditions.

Playback Strategy The informational inputs are merged to form an algorithmic playback strategy for the sound objects, which may include such attributes such as source motion, looping and multitude.

Post Analysis After resynthesis of the textural layer, spectral and temporal information is extracted to ensure the fidelity of the stream and for use in the mixing with the object layer.

3.1.3 Processes

Define & Segment The audio content from the sound database is segmented and defined in terms of the referential taxonomy. The segmentation includes a separation of transient, harmonic and stochastic sources. The content of each segment is decided using the tags associated with the original files and source recognition.

Resynthesis The textural layer is synthesized using concatenative synthesis (mosaicing), wavelet resynthesis and looped playback, with the synthesis parameters provided by the Content Analysis and Textural Parameters analysis blocks.

Pre-playback Conditioning Before mixing with the textural layer, some basic conditioning such as enveloping and filtering are applied to the sound objects to ensure clean playback and spectral separation of the objects.

Mix & Spatialize The sounds are mixed with respect to the playback strategy and textural parameters and spatialized based on the user orientation within soundscape.

3.2 Sound Object Retrieval

From the searches for soundmarks, signals and city information, a set of keywords are obtained for retrieval of sounds from the unstructured database and organized into a sound object list. Using a lexical database search technique, sounds related to each of the items on the sound object list are retrieved for use in the soundscape generation.

3.3 Generation

The purposes of the design described in this section are to create a sense of immersion in the soundscape and to deliver relevant information to the user regarding the environment and conditions. In order to optimize the system for these purposes, the aspects of scene realism, variety (ie. reduction of the perception of loops) and interactivity are given pinnacle importance in the playback strategy and audio processing.

Three forms of generation of sounds are used in this model: sample playback, wavelet resynthesis and concatenative synthesis. Using these different methods allows for optimal source separation and efficiency in processing the differing types of sounds.

3.3.1 Sample Playback

After segmentation and definition, a bank of sound objects for playback is created from which to construct the sound scene. For many sound types such as impulsive events and indicators, which tend to be of relatively short duration, it is most efficient to trigger these sounds directly without analysis and resynthesis. Birchfield et. al. constructed soundscapes using a playback and looping method in a generative soundscape model that adapted to user actions and improved variability by applying probabilities to the sound events [4], using the principles of soundscape composition as described by Barry Truax and R. Murray Schafer.

For interactivity-related sound events, such as impact sounds arising as a result of movement or contact with objects in the scene, sample playback is a natural choice for sound generation. However, the drawback in implementation, in game engines for example, is often the compromise between variety and the amount of memory space to store the samples. Techniques which resynthesize samples to accompany the gestural or contact characteristics, such as modal synthesis, are very useful for transforming the original sound with respect to variations in the interaction [22]. For example, a transformation of the sound of the footsteps to match the ground material would add to the realism aspect, and thereby increase the sense of presence in the environment [6].

Textural sources that have a repeating pattern may also be favored to incorporate a looping playback strategy, in order to reduce the complexity of the application.

3.3.2 Wavelet Analysis and Resynthesis

The analysis of textural sources that are highly stochastic and non-stationary are very suitable for a wavelet decomposition approach [11]. Wavelet basis functions are selected and compared with the signal at different temporal translations with different scalings, to retrieve a set of decomposition coefficients representing the likeness of the signal to the basis functions in time. These coefficients may then be used to resynthesize the texture with time-variance and dynamic changes. Applying a parameterization model depending on the sound type classification, the decomposition coefficients may be used to alter the synthesis result to generate an altered version of the original signal. For example, by varying the time and frequency parameters

of a rain signal decomposition, the output signal can be altered to simulate varying heaviness of rainfall [12].

3.3.3 Concatenative Synthesis

Concatenative synthesis can be thought of as a content-based extension of granular synthesis, where an audio stream is broken into units, which are selected based on likeness to a target signal [15]. This synthesis method is highly suitable for handling sound files from a database, which can be segmented. The size of the units are selected based on the type of source, and the playback of the units is performed using the concatenative synthesis technique to produce a texture from a set of pre-recorded sounds. The grain size should be long enough so that the source of the sound is recognizable, yet short enough to prevent the perception of looping events. For example, the grain size for speech should be long enough to recognize full words in order for the language to be recognizable (ie. $\sim$ 1 second), but not long enough for entire sentences or phrases (ie. $\sim$ 3 seconds) to be discernible to prevent that highly recognizable events are perceived to be repeating during playback. In this way the grain size does not necessarily need to be held constant, and may contain a window of variance in order to prevent periodicity in the artifacts that may arise due to discontinuities between the grains.

In general the choice of grain size can highly affect the resulting texture, and for most naturally recorded sounds a general selection of grain size for a category, such as birds or construction, requires some manual tuning in order to achieve the best result. If a grain size is chosen that is too long in comparison with the event, then a repeating pattern may be perceived and the recognized during subsequent playback. On the other hand, if a grain size is selected that is too short compared with the natural duration of the event, then the naturalness of the signal may be compromised.

As an example, in the case of a recording of construction activity where there may be present both sounds of hammering and sawing, the optimal grain size for the two events would not be equivalent. For the case of a saw, the onset, sustain and decay portions of the event must be present and in the natural order of occurrence, and for hammering the repetition of patterns of more than a few strikes would be recognizable. In both of these cases, the likelihood of distraction for the listener is high, and therefore a variable grain size is necessary for such a recording.

Therefore, for an autonomous concatenative synthesis engine, a segmentation procedure based on the temporal changes in the recording is recommended in order to capture both long events and short events. For this engine, an MFCC-based bayesian information criterion (BIC) segmentation procedure is used to scan a recording for optimal segmentation points [2]. Using an MFCC calculation takes into account the spectral properties of the signal and is based on the Mel-scale, which correlates the frequency spectrum of the signal with the perceptual attribute of timbre. A minimum segmentation length is selected depending on the signal classification, based on Schafer's taxonomy, and the grain sizes longer than that length are chosen according to the optimal segmentation points from the MFCC-BIC calculation. The minimum grain sizes should be selected based on the logic as described above in

the example of speech, where the source should be recognizable for the minimum segment length.

Each sound file before segmentation is pre-conditioned with low-pass and high-pass filtering and normalization. The cutoff frequencies for the filters are decided by the frequency range of the sources within the categories. Normalization ensures that the initial amplitudes of the source categories are equivalent in order to ease the global mixing implementation. After segmentation, the segments' boundaries are cross-faded and their selection during playback is decided according to a euclidean distance calculation with the other segments. Segments whose MFCC coefficients' are outside of the standard deviation are disregarded to remove any extraneous segments from the concatenation.

In a given set of segments, the nearest segments among the set are determined and each consecutive segment is selected randomly from this set[1]. The randomness allows the selection to vary across the entire set, while the selected subset based on the nearest MFCC's ensures that large changes in relative spectral content do not result. In general pre-segmented samples are recommended to be at least 20 times longer than the minimal grain size in order to obtain a sufficient number of samples for concatenation without the perception of loops. These concepts are further discussed in section 5 and continued research in quantifying these requirements based on subjective analyses is recommended.

[1]In implementation, it is recommended that this set be composed of the five nearest segments to allow for variation over long durations of exposure to the source. In deciding this number both the number of segments retrieved from segmentation and the expected duration of exposure to the user must be considered.

Chapter 4

Implementation

As an initial implementation of the autonomous soundscape generation model, a static photorealistic environment was chosen in order to evaluate the use of recorded sounds from unstructured databases in an application that must be consistent with the natural environment. The Google Street View application contains panoramic images from discrete locations within a city, in which the user is free to rotate and can translate to different locations. The sound database used for feeding the soundscape generator was The Freesound Project [1], and as many samples as possible were selected with origins in Barcelona and Spain for purposes of authenticity. For the purposes of this study involving the Street View application, the sound objects are retrieved manually in order to focus on the automonous generation aspect, and the automatic retrieval implementation is reserved as an avenue for future work.

Three locations were selected for simulation and the generated soundscapes were compared with real recordings of the locations through a subjective evaluation.

4.1 Taxonomy

For the initial application of generating soundscapes for Google Street View, a categorization according to Schafer's taxonomy of the potential elements within an outdoor city street context are given in table 4.1.

Table 4.1: Schafer's referential classification applied to the most frequent sources found in a typical city.

Natural	**Human**	**Society**	**Mechanical**	**Quiet**	**Indicators**
Air	Voice	City	Transportation		Horns / Whistles
Birds	Clothing		Construction		Bells / Gongs
Insects	Body		Tools		Telephones
Animals			Industrial		Warning Systems

The categories *Society* and *Quiet* can be disregarded for this study, since *Society* is a general classification for environments and *Quiet* relates to environments with very few or negligible sound sources. The subcategories for each of the other categories must be further broken down and further specified in order to determine

Table 4.2: Elaborated referential classification applied to the most frequent sources found in a typical city.

Natural	Human	Mechanical	Indicators
Wind	Speaking Voice	Cars and Trucks	Car Horns
Birds	Yelling Voice	Motorcycles	Church Bells
Insects	Footsteps	Street Power Tools	Mobile Phones
Dogs	Whistles		Police Sirens
Cats			

the range of possible sources that should be included in the Street View application. Reducing the table to the main four categories, and elaborating on the subcategories, the scene description table for cityscapes for this application is given in table 4.2.

4.2 Generation

4.2.1 Sample Playback

The incorporation of sound objects into the soundscape in the Street View application follows a similar model to that of Birchfield et. al. [4], employing probabilistic playback to objects with probability values based on the location and orientation of the user. The sound objects used in this model include construction noises, indicators (sirens, phones) and soundmarks. For example, if the user has selected a location on a street that is labeled as currently having construction activity, as is provided by the RT Conditions input block (See figure 3.1.), the probability that construction noises will be played back would be high, while if the user were to navigate to another street without construction activity then the probability would lower to zero.

For all of the source layers, the levels of the sources were decreased as the POV of the user was directed away from the location of the source. This strategy is based on the sound design principles and presence factors related to synchrony and multimodal consistency.

4.2.2 Wavelet Analysis and Resynthesis

Wavelet decomposition is very suitable for the Street View application, as a relatively small set of sound recordings may be used to train the model in order to synthesize a manipulable and time-varying texture. This approach is recommended for the textural sources of wind and water as a future improvement to the application, since they are non-stationary and stochastic signals for which a parameterized model for alteration based on an external input may be applied. However it may be argued that for this particular application it is beneficial to preserve the original character of the community-provided sounds, and thus reflect the true quality of the soundscape. If the resynthesized sounds deviate noticeably from the original recordings, then the

application may risk losing its preservational aspect. Therefore, such techniques of resynthesis should not incorporate an abundant amount of parameterization, but instead focus on the aim of preventing the perception of loops during playback.

4.2.3 Concatenative Synthesis

For this application it was decided that the community-provided recordings should be as preserved as possible, in order to reflect the true nature of the urban soundscape as recorded by individuals within the context. Therefore, for textural sources a concatenative synthesis technique is the most suitable method for the application. As described in section 3.3.3 the recordings are segmented according to a MFCC-based BIC technique, where the minimum grain sizes are chosen for the selected categories of the Schafer taxonomy. For this application the grain size selection is performed for the three categories *natural*, *human* and *mechanical*.

As sound recordings of natural elements found in community-driven sound databases for an urban context are primarily composed of city-dwelling birds, the choice of grain size may in theory be selected to be very short, as many bird sounds are impulsive. However, field recordings such as those found in unstructured sound databases contain a combination of bird noises and wind, and therefore it is desirable to retain the long duration of the textural evolutions found in these recordings, and as well to prevent artifacts from the resynthesis discontinuities which may arise due to the stochastic layer in these recordings. Therefore a longer minimum grain size of 4 seconds was selected for the segmentation procedure, however further analysis using many different recordings could result in a more optimal selection. For human sounds, the minimum grain size was selected based on the logic described in section 3.3.3 to be one second, and for construction two seconds to prevent the recognition of long patterns of impact noises.

Table 4.3 shows the selected minimum grain sizes and cutoff frequencies for the source categories.

Table 4.3: Selected cutoff frequencies and minimum grain sizes for the source categories.

	Natural	**Human**	**Mechanical**
HP Cutoff Frequency (Hz)	300	200	50
LP Cutoff Frequency (kHz)	13	10	10
Min Grain Size (s)	4	1	2

4.3 System Architecture

The Street View application is itself a Flash object, and thus the interaction with the object was created in Actionscript 3.0 and communicated with the Street View object through Javascript. The sound generation was initially performed on a host machine with Supercollider, and received messages from the Flash object for triggering the concatenation and playback via OSC messages over the server. While

Supercollider is a highly flexible choice for sound generation, and allows for the extension into low level processing, for this application it was decided to move the sound generation to Actionscript in order to bundle the processes. Furthermore this architecture allows for playback to be generated on the client-side, and eases the streaming requirements if the system were to be implemented on a larger scale for community usage or experimentation. As the low level processing and segmentation is performed off-line, only the playback strategy, spatialization and concatenation timing is performed on the client-side, and therefore Actionscript is applicable for sound generation for the Street View application.

A screenshot of the simulation interface is shown in figure 4.1, where the sources are shown in the bird's eye view map at their locations relative to the user on the right side with circles representing their current amplitudes.

The segments and their information regarding MFCC selection criteria (in .mp3 and .txt format, respectively), are retrieved by the Flash object and the location, point-of-view (POV) and time of day are the logical inputs to the system. The following subsections describe the playback strategies for the three source categories.

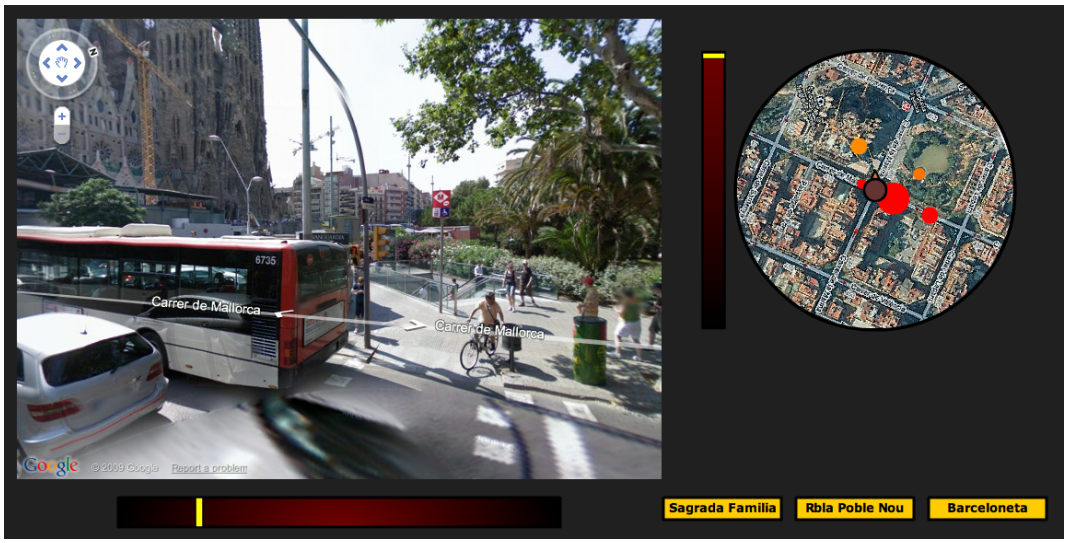

Figure 4.1: Screenshot of the Street View implementation.

4.3.1 Natural

The chosen natural sources contain recordings of urban-dwelling birds and water. The bird sounds are divided into the three categories *trees*, *street* and *beach*. The logical selection of the category depends upon location, where parks and areas with dense forestation result in samples from the *trees* category, urban street locations *street* and of course proximity to the sea increases the likelihood of selecting from the *beach* category. The placement of these sources relative to the user's POV is consistent with the imagery, as the sources are placed on the map locations for which they relate. Recordings of between 60 and 120 seconds were used to create between 10 and 20 segments from each recording, and a low-pass filter cutoff frequency of 13

kHz and a high-pass filter cutoff frequency of 300 Hz was chosen based on the likely frequency range of content for the natural sources.

4.3.2 Human

Recordings of voices were selected and categorized based on density and time of day. Locations of heavy pedestrian traffic result in more source layers, and the time of day determines the density of voices within the samples. Samples of between 30 and 120 seconds were used and segmented using a minimum segment length of one second. The low-pass filter cutoff frequency used was 10 kHz and the high-pass cutoff was 200 Hz, in accordance with the typical frequency range of speech.

4.3.3 Machinery

Machinery sounds included both construction and traffic noises, as well as signals such as sirens and horns. The construction sounds were concatenated using a minimum grain size of one second, with a low-pass cutoff frequency of 10 kHz and a high-pass cutoff frequency of 50 Hz.

Traffic sounds were synthesized using a sample playback method, with individual samples of vehicles passing. In some applications traffic may be considered a textural source, however as the user is positioned in the street, the passing vehicles are highly active in this context (see section 2.2.3). In addition, layering these sounds algorithmically allows for much more control of the conditions, and alleviates complications with segmentation and looping of field recordings of many vehicles simultaneously. The segments were divided into two general categories of *slow* and *fast*, and subcategories of *light*, *medium*, *heavy*, *truck* and *motorcycle*. The traffic conditions are input to the system based on the location and time of day, and dictate which and how many samples are used and the timing between their playback. For this simulation only the situation of passing traffic was assumed, although the model should be extended to include idling, decelerating and accelerating traffic as well.

As traffic in urban areas tends to be grouped due to clustering at intersections, the simulation uses a clustering technique where sets of vehicles are layered with some separation between the clusters, as shown in figure 4.2. The group separation is decided by the time of day input, and is discretized into low, medium and high, where the minimum group separation times are 8, 12 and 20 seconds, respectively. A random fluctuation variable is added to these values with a range of half of the group separation time. Linking this variable to the group separation time simulates the increasing sporadicity of traffic with less density.

The elements within the group are composed of samples of the individual vehicles for the categories discussed above, and each element is selected randomly during playback. In this manner, the density of a certain type of vehicle, *light* for example, is decided by the number of samples in the sound bank relative to the other subcategories. The number of elements in each group is determined by an input based on the locality, which is the number of lanes in the road at the current location. There is as well a random fluctuation variable for the time offset of the elements

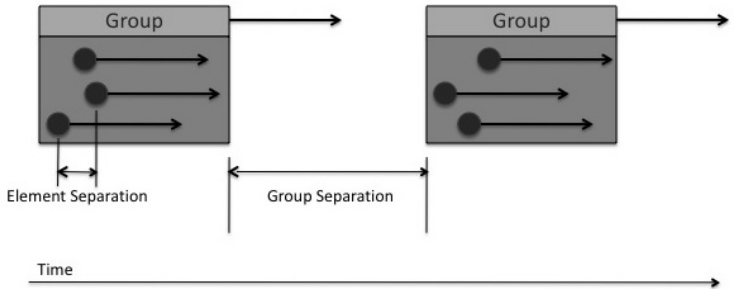

Figure 4.2: Diagram representing the generalized functionality of the traffic simulation for the Street View application.

within the group relative to one another, which is within the range of the group segmentation time divided by eight. The elements are individually spatialized in relation to the user position to give the impression of direction, which is correlated with the direction of traffic on the street.

4.4 Evaluation

4.4.1 Criteria

In the Street View application, the factors related to control and the visual percept are quite limited, since the visual image to which the user is subjected is static. The possible interactions with the environment are location selection and POV alteration. As the soundscape represents a dynamic image, while the image presented to the user is static, it is apparent that the attainable degree of immersion is much lower than a typical virtual environment, for which Singer's parameters are meant to evaluate. Therefore, many of the parameters are not used in this evaluation as they are not highly applicable to interaction within a VE composed of a static image and a dynamic soundscape.

The factors from table 2.1 that are determined to be most relevant for the evaluation were selected, and a description of each as given by Singer et. al., is provided in table 4.4. These factors are to be evaluated using the procedure described in the following section.

4.4.2 Procedure

As presence is a psychological phenomenon, the manner of its measurement and quantification is often disputed. The two main categories measuring presence are subjective and objective techniques. Subjective techniques require the subject to verbalize or express their perceived sense of presence through the use of questionnaires or quizzes [17], [16], [18], and objective methods observe behavioral [9] and physiological responses of the subject during the experiment [3]. For this study, a subjective analysis is used to evaluate the system for this stage of development of the soundscape generation model.

Table 4.4: Selected parameters and their definitions from Singer et. al. [17] used for evaluation of the Street View sonification.

Anticipation of Events	Individuals probably will experience a greater sense of presence in an environment if they are able to anticipate or predict what will happen next, whether or not it is under personal control.
Physical Environment Modifiability	Presence should increase as ones ability to modify physical objects in that environment increases.
Environmental Richness	The greater the extent of sensory information transmitted to appropriate senses of the observer, the stronger the sense of presence will be.
Consistency of Multimodal Information	The information received through all modalities should describe the same objective world.
Degree of Movement Perception	Presence can be enhanced if the observer perceives self-movement through the VE, and to the extent that objects appear to move relative to the observer.
Active Search	An environment should enhance presence when it permits observers to control the relation of their senses to the environment.To the extent that observers can modify their viewpoint to change what they see, or to reposition their head to affect binaural hearing, or to search the environment haptically, they should experience more presence.
Scene Realism	Scene realism ... refers to the connectedness and continuity of the stimuli being experienced (as governed by scene content, texture, resolution, light sources, field of view (FOV), dimensionality, etc.).
Information Consistent with Objective World	The more consistent the information conveyed by a VE is with that learned through real-world experience, the more presence should be experienced in that VE.

Figure 4.3: User interface for subjective evaluation.

The analysis was performed using eight participants, who were presented soundscapes generated using the Street View sonifier and recorded soundscapes for three locations using headphones (Sennheisser HD650). The recorded soundscapes were quadraphonic recordings situated at each of the locations between 12 and 2 P.M. on weekdays with a Zoom H2 handheld recorder. The four channels were used to create a spatialized reproduction in Actionscript with mixing orientation linked to user orientation. Each recorded soundscape was one minute in duration and selected manually beforehand from a longer recording, with the intent to include various sound sources without heavy weighting on any one sound object in particular.

Table 4.5: Samples and settings for subjective evaluation.

	Sagrada Familia	**Rambla Poble Nou**	**Barceloneta**
Birds - Street	$Level = 20\%$, $Pan = 75°$	$Level = 10\%$, $Pan = -45°$	$Level = 10\%$, $Pan = 60°$
Birds - Trees	$Level = 60\%$, $Pan = -45°$	$Level = 40\%$, $Pan = -135°$	$Level = 20\%$, $Pan = 45°$
Birds - Beach	—	—	$Level = 10\%$, $Pan = 110°$
Voices - Medium	—	$Level = 30\%$, $Pan = 90°$	$Level = 20\%$, $Pan = -120°$
Voices - Crowd	$Level = 80\%$, $Pan = 90°$	$Level = 80\%$, $Pan = -110°$	$Level = 60\%$, $Pan = 160°$
Construction	$Level = 100\%$, $Pan = -135°$	—	$Level = 40\%$, $Pan = -100°$
Traffic	$Lanes = 4$, $Density = "high"$	$Lanes = 1$, $Density = "mid"$	$Lanes = 2$, $Density = "mid"$

The generated soundscapes included samples from the three Schafer categories *Natural*, *Human* and *Machine*, and traffic simulation settings representative of the number of lanes and time of day. The same sound samples were used for each location with variations in level and spatialization manually determined to represent the locality[1]. A description of the samples used in each location, their spatializations

[1]Samples not referentially relating to a locality were not included in the generated soundscape, as is shown in table 4.5.

Table 4.6: Second phase of subjective evaluation - Questionnaire.

1.	How well could you identify individual sound sources in the soundscape?
2.	How well could you actively localize individual sound sources in the soundscape?
3.	How compelling was your sense of objects moving through space?
4.	How much did the auditory aspects of the environment involve you?
5.	How compelling was your sense of movement (turning) inside the virtual environment?
6.	How much did your experiences in the virtual environment seem consistent with your real-world experiences?

relative to the user's initial orientation and their levels relative to their normalized maxima are given in table 4.5, along with the traffic simulation settings.

The subject was situated in front of a computer with the application running, which contained the panoramic image and a control application for changing the POV. A screenshot of the user interface is shown in figure 4.3. The user was asked to change the POV for each of the three locations, and was then given a questionnaire afterwards as is shown in table 4.6, which was composed of the questions in the Singer et. al experiments that are related to the chosen factors from table 4.4. For each of the questions, the user selected a box within a range as is shown in table 4.7.

Table 4.7: Subject response example box.

Not at all		Moderately		Very well
o	o	o	o	o

The users were instructed to randomly select their first location, and use the panning bar underneath the Street View window to explore the image and auditory environment. They were given the questions beforehand and allowed to answer them during the experiment while changing locations and soundscapes. The generated and recorded soundscapes were not placed in any particular order and the users were not aware before the experiment that some soundscapes were recorded and some were generated. Two of the recorded soundscapes were in region B and one was in region A, where the regions can be seen figure 4.3. The users were not given any time limitation and were allowed to spend as little or as much time as they desired at each location.

4.4.3 Results and Discussion

The results of the analysis showed that the subjects generally rated the generated soundscapes higher than the recorded soundscapes for each of the individual questions, by a margin of $15 - 20\%$. While there were not enough subjects for drawing statistically significant conclusions, the results demonstrate that the generated soundscapes are at least acceptable in comparison with actual recordings. It is especially interesting to note that for the question related to the consistency with real-world experiences (see table 4.6), the generated soundscapes actually scored higher than actual recordings. This result is consistent with the experimenter's observation that none of the subjects were able to identify which soundscapes were real and which were generated following the experiment, when asked informally if they noticed a distinctive difference.

The amount of favorability related to the individual questions did not show any one attribute to a be prominent contributor to the results, and the distribution was even with respect to the questions related to sensory factors and realism factors (see appendix B). Furthermore, the three locations returned similar results with respect to the difference between generated and recorded soundscapes. The normalized responses to the individual questions for the three locations are shown in figure 4.4.

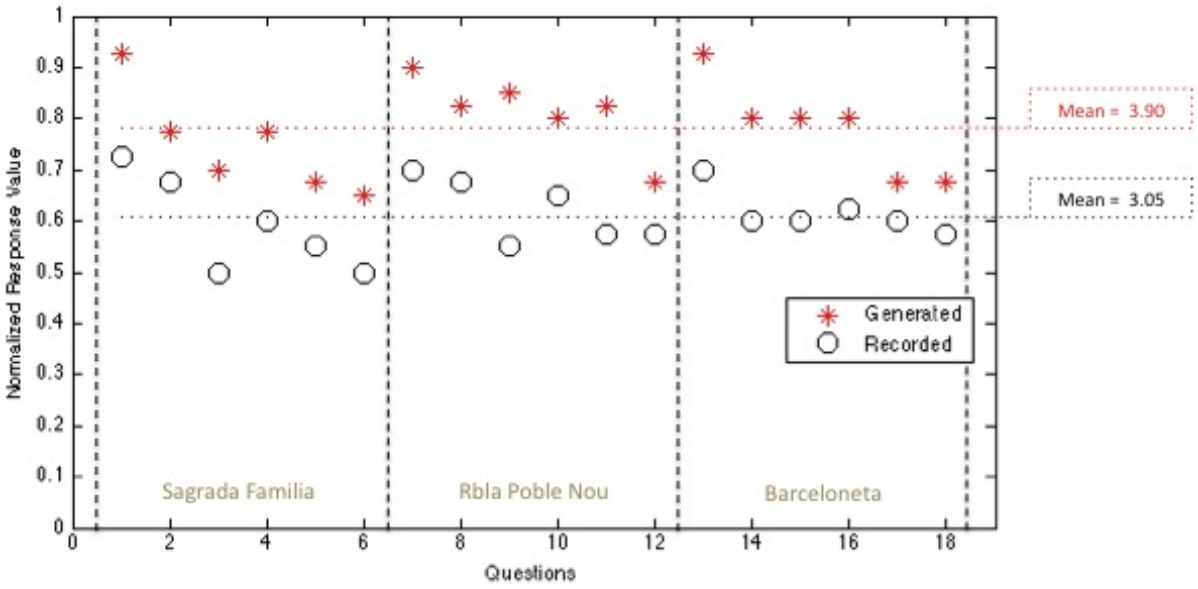

Figure 4.4: Normalized responses averaged over all participants in the subjective evaluation for each question.

From the statistical results and the comments of the participants following the experiment, some explanations for the favorability of the generated soundscapes over the recorded soundscapes may be discussed. As many of the questions were related to the identification of sources, localization and dynamism of sound objects, the spatialization and mixing of the separate layers seem to have had a pronounced effect. The participants were more easily able to distinguish and localize objects in the sound scene and perceived a higher degree of movement of objects with the gen-

erated soundscapes using separated source layers and individual layer spatialization. Although the recorded soundscapes were created with quadraphonic recordings and spatialized relative to user orientation, it is recommended for future evaluations to use ambisonic or head-related transfer functions to establish a more realistic spatialization for comparison with the generation model. Furthermore, some users noted that the traffic seemed more congruent with their positions for the generated soundscapes, which is logical since the recordings were made beside the road as opposed to within the flow of traffic. This observation however demonstrates the effect of multi-modal consistency on presence for this application.

The initial levels of each of the source layers was set manually prior to the experiment, in order to create a suitable mixing of the source layers. The samples were normalized prior to segmentation and playback, and it was found during development of the application that level adjustments were necessary and in some cases required very drastic reductions. In order to automate the level mixing process, a perceptual quantification of the source layers' levels must be incorporated for future implementations.

It was also noted by some participants that the recognition of sources related to the location was a large factor in their feeling of being situated within the environment, such as seagulls near the beach. In the recorded soundscapes the natural sounds of the location were not perceived to be as representative for the location as those selected for the generated soundscapes. Selecting sources particular for a location may exaggerate the actual sonic environment creating an augmented sonic reality for the image, and thereby increase the amount of sensory information available to the user.

The design methodology applied to this application seems to have enhanced the immersive experience for the participants as compared with actual recordings due to improved source identification and location, and multi-modal consistency.

Chapter 5

Discussion and Conclusions

The results of this study establish autonomous generation of soundscapes using content from unstructured sound databases as a viable avenue for enhancing the sonic environments of VE's. The combination of sound design principles from the fields of acoustic ecology and media seems as well to lay a groundwork for autonomous soundscape generation, as the results of the evaluation demonstrated a favorable response compared to actual soundscape recordings. However, with the key components of retrieval, sound object selection and initial level mixing performed manually for the simulation, it also exhibits that the implementation of the model must be further developed in order to reach fully autonomous functionality.

The use of automatic level mixing with user orientation, in addition to spatialization, seemed to result in an improved sense of space for the participants, and better localization and identification of individual sources. Additionally, the users were not able to distinguish the generated soundscapes from the natural recordings, which suggests that the concatenative synthesis methodology did not cause perceptual anomalies in the generation, and that the use of layered recordings from an unstructured database can create the illusion of a natural soundscape.

The concatenative synthesis method employed in this model, based on BIC segmentation and MFCC proximity, was observed during development to contain some possibly perceivable instabilities in the samples with notable stochastic components. Although the MFCC selection criterion alleviates these instabilities to some extent, for a single source layer without other layers to mask the spectrally varying stochastic component, these instabilities may contribute to a perceptually observable discontinuity in the audio stream. Further research regarding the masking criteria necessary for ensuring that these variances in spectral content are not perceptually relevant are recommended for future work. The generation model proposed in chapter 3 recommends the use of wavelet resythesis for the stochastic source layers, and its incorporation into a future implementation may resolve this issue.

Selection of the minimum grain sizes for the individual source layers was decided based on the length of the shortest audio event that was desired for capture. This method proposes to use a minimum grain size for each of the sources in Schafer's taxonomy, which are chosen for a particular application, and for the implementation in the Street View sonification the method was determined to be successful. Further analysis using many different samples may demonstrate however that the selection

may be further optimized for each particular recording using a procedure based on detected event lengths within the recording.

The subjective evaluation insisted that the positioning and POV of the user relative to the sources are highly affective attributes of the soundscape, which are believed to be the primarily contributing factors in the favorability of the generated soundscapes over the recorded soundscapes. The correlation of the user's POV and the location of sound objects in the sound scene was used to adjust the levels and of course the spatialization of the source layers, but may be further enhanced with the use of audio effects such as dynamic filtering and compression. Such optimizations would likely further improve the multi-modal consistency and source identification properties related to the sense of presence.

In chapter 3 the concept of the use of metadata related to user interactivity and trends was discussed as an input to the generative strategy. The use of metadata is highly dependent on the application, but carries the potential for enrichment of the playback and generation strategies that may not be available based on analysis of the audio content alone. For the Street View application for example, metadata related to user feedback regarding the relative levels of sources could alleviate the difficulties discussed in section 4.4.3 with automatic mixing. Such user-driven feedback would ensure the most comfortable and desirable mix based not only on loudness perception but also on the relative focal draw of the individual source layers.

The implementation of the model in a photorealistic application demonstrated that it is acceptable for creating realistic sonic environments, however the model lacks validation for a less realistic VE where the use of naturally recorded sound textures as found in unstructured databases may not be as successful. It is expected that the selection of audio content must be suitable for the context of the application, and therefore the model should be validated for non-photorealistic applications, from which a methodology for retrieval based on application context may be derived.

The use of a photorealistic application did however expose the compromise between realism and emotional induction in soundscape design. Although realism may be considered to be of primary importance for the Street View sonification, it was found that users responded favorably to the use of sounds that exaggerated the natural components of the soundscape for the image. The sound objects particular for a location were accentuated through level and density parameters, and in turn contributed to a general preference over the actual recordings, in which these sources were not as prominent. The balance between realism and emotional induction is a design principle that should be evaluated for each application. Further study is warranted for quantifying this balance with relation to the soundscape generation model. A further extension of this notion is the incorporation of diegetic sounds into the model, as discussed in section 2.2. For the Street View application for example, it may be conceived to incorporate narration or instructional sounds to give the user information regarding the location, if so desired by the user. These additions may contribute to the informational exchange between the application and the user and thereby increase its beneficial value.

Appendix A

R. Murray Schafer's Taxonomy

1. **Natural Sounds**
    - Sounds of creation
    - Sounds of apocalypse
    - Sounds of water
    - Sounds of air
    - Sounds of earth
    - Sounds of fire
    - Sounds of birds
    - Sounds of animals
    - Sounds of insects
    - Sounds of fish and sea creatures
    - Sounds of seasons
2. **Human sounds**
    - Sounds of the voice
    - Sounds of the body
    - Sounds of clothing
3. **Sounds and society**
    - General description of rural soundscape
    - Town soundscapes
    - City soundscapes
    - Maritime soundscapes
    - Domestic soundscapes
    - Sounds of trades, professions and livelihoods
    - Sounds of factories and offices
    - Sounds of entertainment
    - Music
    - Ceremonies and festivals
    - Parks and gardens
    - Religious festivals
4. **Mechanical Sounds**
    - Machines
    - Industrial and factory equipment
    - Transportation machines
    - Warfare machines
    - Trains and trolleys
    - Internal combustion engines

- Aircraft
- Construction and demolition equipment
- Mechanical tools
- Ventilations and air-conditioners
- Instruments of war and destruction
- Farm machinery

5. **Quiet and Silence**
6. **Sounds as Indicators**

- Bells and gongs
- Horns and whistles
- Sounds of time
- Telephones
- Warning systems
- Signals of pleasure
- Indicators of future occurrences

Appendix B

Evaluation Results

The following pages contain the raw results and calculations of the subjective evaluation on which the discussion and conclusions in sections 4.4.3 and 5 are based.

Questions: *scale 1 → 5*

Q1: How well could you identify individual sound sources in the soundscape?

Q2: How well could you localize individual sound sources in the soundscape?

Q3: How compelling was your sense of objects moving through space?

Q4: How much did the auditory aspects of the environment involve you?

Q5: How compelling was your sense of movement (turning) inside the virtual environment?

Q6: How much did your experiences in the virtual environment seem consistent with your real-world experiences?

Generated Soundscape Responses to Individual Questions

	Location 2						**Location 2**						**Location 3**					
	Q1	Q2	Q3	Q4	Q5	Q6	Q1	Q2	Q3	Q4	Q5	Q6	Q1	Q2	Q3	Q4	Q5	Q6
Subject 1	5	4	4	5	5	4	4	5	5	4	5	2	5	3	4	3	3	3
Subject 2	5	5	3	4	4	3	5	5	4	5	4	3	5	5	5	NONE	4	4
Subject 3	5	4	3	4	3	3	4	4	3	4	3	3	5	5	3	5	2	4
Subject 4	5	3	4	4	4	3	5	4	5	4	5	4	5	4	4	4	5	4
Subject 5	5	4	4	5	4	4	5	5	5	4	4	5	5	5	5	5	4	5
Subject 6	4	3	4	4	4	3	5	4	4	5	5	5	4	4	4	4	4	3
Subject 7	3	3	3	2	2	3	3	3	4	3	3	3	4	2	3	2	2	2
Subject 8	5	5	3	3	1	3	5	3	4	3	4	2	4	4	4	5	3	2
MEAN	4.63	3.88	3.50	3.88	3.38	3.25	4.50	4.13	4.25	4.00	4.13	3.38	4.63	4.00	4.00	4.00	3.38	3.38

Recorded Soundscape Responses to Individual Questions

	Location 2						**Location 2**						**Location 3**					
	Q1	Q2	Q3	Q4	Q5	Q6	Q1	Q2	Q3	Q4	Q5	Q6	Q1	Q2	Q3	Q4	Q5	Q6
Subject 1	3	2	2	NONE	1	2	3	2	2	3	2	2	4	4	4	3	4	4
Subject 2	4	4	3	3	4	2	4	4	3	3	3	4	5	3	2	3	2	3
Subject 3	4	4	4	3	3	3	4	4	3	3	2	3	4	4	3	4	3	3
Subject 4	2	2	1	3	3	2	5	4	4	4	5	4	1	2	2	1	3	1
Subject 5	4	4	3	3	4	4	3	3	3	3	3	3	4	3	4	3	4	4
Subject 6	5	5	4	5	4	4	2	3	3	3	4	3	4	4	4	4	4	4
Subject 7	4	2	2	2	2	1	3	2	3	4	3	2	3	2	3	4	2	2
Subject 8	3	4	1	2	1	2	4	5	1	3	1	2	3	2	2	3	2	2
MEAN	3.63	3.38	2.50	3.00	2.75	2.50	3.50	3.38	2.75	3.25	2.88	2.88	3.50	3.00	3.00	3.13	3.00	2.88

Average Responses to Individual Questions

		Location 1		Location 2		Location 3		Mean	
Questions	Factors	Generated	Recorded	Generated	Recorded	Generated	Recorded	Generated	Recorded
Q1	RF, SF	4.63	3.63	4.50	3.50	4.63	3.50	4.58	3.54
Q2	RF, SF	3.88	3.38	4.13	3.38	4.00	3.00	4.00	3.25
Q3	SF	3.50	2.50	4.25	2.75	4.00	3.00	3.92	2.75
Q4	SF	3.88	3.00	4.00	3.25	4.00	3.13	3.96	3.13
Q5	SF	3.38	2.75	4.13	2.88	3.38	3.00	3.63	2.88
Q6	RF, CF	3.25	2.50	3.38	2.88	3.38	2.88	3.33	2.75
Mean		3.75	2.96	4.06	3.10	3.90	3.08	3.90	3.05

Normalized Differences (Generated - Recorded)

	Location 1	**Location 2**	**Location 3**	**Mean**
Q1	0.20	0.20	0.23	0.21
Q2	0.10	0.15	0.20	0.15
Q3	0.20	0.30	0.20	0.23
Q4	0.18	0.15	0.18	0.17
Q5	0.13	0.25	0.08	0.15
Q6	0.15	0.10	0.10	0.12
Mean	0.16	0.19	0.16	0.17
Sensory Factors	0.16	0.21	0.18	0.18
Realism Factors	0.15	0.15	0.18	0.16
Control Factors	0.15	0.10	0.10	0.12

Bibliography

[1] Freesound.org. www.freesound.org.

[2] X. Anguera and J. Hernando. Xbic: Real-time cross probabilities measure for speaker segmentation. *Univ. California Berkeley, ICSIBerkeley Tech. Rep*, 2005.

[3] W. Barfield and C. Hendrix. Presence and performance within virtual environments. *Presence: Teleoperators & Virtual Environments*, 3, 1996.

[4] D. Birchfield, N. Mattar, and H. Sundaram. Design of a generative model for soundscape creation. In *International Computer Music Conference, Barcelona, Spain*, 2005.

[5] Michel Chion, Claudia Gorbman, and Walter Murch. *Audio-Vision*. Columbia University Press, April 1994.

[6] P. Cook. Modeling bills gait: Analysis and parametric synthesis of walking sounds. *Proc. Audio Engr. Society*, 22, 2002.

[7] P. R. Cook. Toward physically-informed parametric synthesis of sound effects. In *1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, page 15, 1999.

[8] C. Fellbaum et al. *WordNet: An electronic lexical database*. MIT press Cambridge, MA, 1998.

[9] Jonathan Freeman, S. E. Avons, Don E. Pearson, and Wijnand A. IJsselsteijn. Effects of sensory information and prior experience on direct subjective ratings of presence. *Presence: Teleoperators & Virtual Environments*, 8(1):1–13, February 1999.

[10] W. W. Gaver. How do we hear in the world? explorations in ecological acoustics. *Ecological Psychology*, 5(4):285–313, 1993.

[11] N. E. Miner and T. P. Caudell. A wavelet synthesis technique for creating realistic virtual environment sounds. *Presence: Teleoperators & Virtual Environments*, 11(5):493–507, 2002.

[12] N. E. Miner and T. P. Caudell. Using wavelets to synthesize stochastic-based sounds for immersive virtual environments. *ACM Transactions on Applied Perception (TAP)*, 2(4):521–528, 2005.

[13] Davide Rocchesso and Federico Fontana. *The sounding object*. Mondo Estremo, 2003.

[14] R. Murray Schafer. *The Soundscape: Our Sonic Environment and the Tuning of the World*. Destiny Books, 1994.

[15] D. Schwarz. A system for data-driven concatenative sound synthesis. In *Proc. Cost- G6 Conf. on Digital Audio Effects (DAFX)*, pages 97–102, Verona, Italy, 2000.

[16] S. Serafin. Sound design to enhance presence in photorealistic virtual reality. In *Proceedings of the 2004 International Conference on Auditory Display*, pages 6–9, 2004.

[17] Michael J Singer and Bob G. Witmer. Measuring presence in virtual environments: A presence questionnaire. *PRESENCE*, 7:225–240, 1998.

[18] M. Slater and M. Usoh. Representations systems, perceptual position, and presence in immersive virtual environments. *Presence: Teleoperators and virtual environments*, 2(3):221–233, 1993.

[19] D. Sonnenschein. *Sound design: The expressive power of music, voice and sound effects in cinema*. Michael Wiese Productions, 2001.

[20] P. Turner, I. McGregor, S. Turner, and F. Carroll. Evaluating soundscapes as a means of creating a sense of place. In *International Conference on Auditory Display*, 2003.

[21] Andrea Valle, Mattia Schirosa, and Vincenzo Lombardo. A framework for soundscape analysis and re-synthesis. In *Proceedings of the SMC 2009*, Porto, Portugal, July 2009.

[22] Kees van den Doel, Paul G. Kry, and Dinesh K. Pai. FoleyAutomatic: physically-based sound effects for interactive simulation and animation. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 537–544. ACM, 2001.